Published in final edited form as:
Psychoneuroendocrinology. 2014 November ; 49: 299–309. doi:10.1016/j.psyneuen.2014.07.022.

Variability and reliability of diurnal cortisol in younger and older adults: Implications for design decisions

Suzanne C. Segerstrom, Ph.D.,
Department of Psychology, University of Kentucky

Ian A. Boggero, M.S.,
Department of Psychology, University of Kentucky

Gregory T. Smith, Ph.D., and
Department of Psychology, University of Kentucky

Sandra E. Sephton, Ph.D.
Department of Psychological and Brain Sciences, University of Louisville

Abstract

The extant research is inconclusive regarding the best sampling methods to construct reliable measures of between-person differences in derived parameters of diurnal cortisol, and no study provides such recommendations for detecting within-person changes. These studies determined how many days of sampling are necessary to assess between-person differences and within-person changes over multiple occasions in diurnal mean, diurnal slope, and area under the curve (AUC). Generalizability and decision analyses were conducted on diurnal salivary cortisol data from two separate longitudinal studies, one with younger adults (N = 124) and one with older adults (N = 148). In both studies, results indicated that 3 days of data collection provided the minimal level of reliability in mean cortisol to detect between-person differences; 4–8 days were necessary to reliably assess AUC, and 10 days for cortisol slope. Similarly, in order to reliably characterize within-person changes across occasions, at least 3 days of data collection were needed for mean cortisol and AUC and 5–8 days for slope. Results also indicated that only two samples per day, taken morning and evening, could faithfully reproduce the diurnal slope calculated from 3 or 4 samples (r = .97–.99). Instead of having participants provide many samples per day over the course of a few days, we recommend collecting fewer samples per day over more days.

Address correspondence to: Suzanne C. Segerstrom, Ph.D., University of Kentucky, Department of Psychology, 125 Kastle Hall, Lexington, KY 40506-0044, Phone 859-257-4549, FAX 859-323-1979, segerstrom@uky.edu.

CONTRIBUTORS

Suzanne C. Segerstrom was the PI of both projects, supervised all aspects of data collection, performed the data analysis, and wrote the manuscript with Mr. Boggero.

Ian A. Boggero and Gregory T. Smith wrote the manuscript with Dr. Segerstrom.

Sandra E. Sephton performed the analyses of cortisol data including assay and cleaning and provided feedback in the writing of the manuscript.

Keywords

salivary cortisol; generalizability theory; reliability; longitudinal design; individual differences

Many studies explore the relationship between salivary cortisol and personal, situational, and environmental psychosocial variables. The present paper generates “physiometric” (Segerstrom & Smith, 2012) data to inform design decisions about diurnal cortisol in such studies. Measures taken on multiple people at multiple times contain three sources of variability: between-person, within-person, and measurement error. Between-person variability reflects how one person differs from another. Within-person variability reflects how people differ from themselves across time. Measurement error reflects the difference between the true state of a person at the time of measurement and the results of that measurement. We provide estimates of these sources of variability and use these estimates to predict generalizability associated with various designs. By doing so, we hope to promote (1) study design that respects the measurement properties of diurnal cortisol, (2) the regular reporting of physiometric information, and (3) further research specifically testing generalizability in varied samples and varied designs and over varied intervals. Understanding and maximizing the reliability of measures of biological variables is important to study designs that will yield accurate estimates. In a Monte Carlo analysis, when predicting an unreliable (.29) measure of immunity taken from a single occasion of measurement, 227/1000 beta weights fell outside the 95% confidence interval obtained using a reliable (.84) measure taken from an aggregation across occasions. Aggregating across more measurement occasions and thereby increasing reliability decreased the number of anomalous results (Segerstrom, Lubach, & Coe, 2006).

When assessed on consecutive days, approximately half of the variance in both cortisol level and diurnal slope is stable between-person variability, and half is idiosyncratic to the day (Kirschbaum et al., 1990; Kraemer et al., 2006; Golden et al., 2011; Kertes & van Dulmen, 2012; though see Hruschka et al., 2005). However, as intervals increase to weeks or months, the proportion of stable between-person variability decreases to approximately 10% (Kirschbaum et al., 1990; Hruschka et al., 2005; Rotenberg et al., 2012; Shirtcliff et al., 2012; c.f., Gex-Fabry et al., 2012). A person’s cortisol parameters measured today have limited generalizability to other time points, even yesterday or tomorrow. The necessity of multiple assessment days in extracting stable person variance in cortisol parameters is therefore widely recognized; however, recommendations for how many days vary. The MacArthur Network’s online recommendation is “3–4 days to get a reliable assessment of a ‘trait’ daily concentration (area-under-the-curve), and 6 or more days to get a reliable assessment of a ‘trait’ rhythm.” Kraemer and colleagues (2006) estimated 2–3 days to reliably estimate “trait” slope. Hruschka and colleagues (2005) estimated 14–22 days to reliably estimate “trait” slope (albeit fewer for cortisol level). Clearly, “more studies need to be carried out … to define the precise parameters of sampling requirements” (Goodyer et al., 2001, p. 243).

In studies with multiple measurement occasions (e.g., an intervention study with baseline, post-intervention, and follow-up or a longitudinal study with annual measurement

occasions), cortisol might be measured on multiple days within each occasion. Variance estimates for single-day measurements are of limited utility for design decisions involving multiple measurements at multiple time points (Kirschbaum et al., 1990; Hruschka et al., 2005; Rotenberg et al., 2012; Shirtcliff et al., 2012). Instead, decisions may focus on the number of days and occasions required for measures to discriminate people from each other at the same occasion, different occasions, or across an aggregate of occasions with adequate reliability. Furthermore, there is the question of how many days per occasion would be required for a measure to discriminate a person at one occasion from him- or herself a different occasion with adequate reliability (e.g., before and after treatment) (Cranford et al., 2006). Table 1 provides examples of study designs that focus on discriminating people from each other when measured once at the same occasion (I), when measured once at different occasions (II), and when measured across several occasions (III), as well as a study design that focuses on discriminating change within people across several occasions (IV).

The challenge in designing diurnal cortisol studies is to maximize variability arising from the facet of interest (e.g., differences between or changes within people) and minimize variability due to other facets and measurement error. Classical test theory assumes that any observed score is the result of a true score and error variance. Generalizability theory (Shavelson & Webb, 1991; Brennan, 2001) extends this assumption to encompass multiple sources of variance, for example, due to people, occasions, and their interaction. What variance is of interest and therefore “true score”, however, depends on the design and research question. For example, to demonstrate that “individuals differ in their patterns of cortisol secretion and … these differences exhibit some stability over time” (Hruschka et al., 2005, p. 699), the “true score” of a cortisol parameter would consist of person variance, and “error” variance would arise from day and occasion variance, as well as their interactions (Cranford et al., 2006). In other cases, the variance across occasions (i.e., change over time) may be the “true score”, and “error” variance would arise from variance among the days comprising each occasion. (Measurement error per se typically accounts for very little variability in cortisol when assays are done competently; Kirschbaum et al., 1990; Kertes & van Dulmen, 2012; Marceau et al., 2013). Even the variability of scores can be the “true score” of interest (e.g., Marceau et al., 2013), but it is still necessary to provide for adequate reliability in the measurement of variability, itself a methodological challenge (Estabrook, Grimm, & Bowles, 2012).

Finally, derived measures such as diurnal slope require consideration of another design decision, involving the number of samples collected each day and their timing (Kudielka et al., 2012). Some studies of diurnal slope have asked participants to provide over 40 samples (Ice et al., 2004), but other evidence suggests two to three samples per day may effectively reproduce the slope calculated from more samples (Kraemer et al., 2006). Collecting many samples per day is expensive for researchers and burdensome to participants. Therefore, further research is needed to clarify how well fewer samples reproduce slopes calculated using more samples per day.

The Current Studies

We applied generalizability theory (Shavelson & Webb, 1991; Brennan, 2001) to two longitudinal studies, one with younger adults and one with older adults. Generalizability theory (Shavelson & Webb, 1991; Brennan, 2001) is an extension of reliability theory with the capacity to estimate what percentage of a single value or assessment is due to stable individual differences, measurement occasions, or interactions between individuals and occasions. A generalizability (G) study uses variance estimates (such as those derived when conducting ANOVA or similar models) to partition variance among these different components. A decision (D) study then uses the results of the G study to inform design decisions that have the goal of achieving adequate measurement reliability and generalizability. The analyses therefore described the sources of variance in diurnal cortisol and provided design decision parameters.

The main design analysis addressed how many days of sampling would be necessary to characterize between-person differences and within-person changes in diurnal cortisol measures with adequate reliability. We also tested whether the data conformed to indications that few samples, particularly if taken at waking and late in the day, can reproduce the diurnal slope calculated with more samples with good fidelity (Kraemer et al., 2006). Finally, we demonstrate that "noise" from one analysis perspective (i.e., daily variability that comprises error in the analysis of between-person differences) can be substantive in another analysis perspective (i.e., daily and occasion variability can be predicted from within-person changes in affect and health behavior).

Study 1

First-year law students collected saliva samples at 5 times over 3 consecutive weekdays at 5 different occasions. Students were assessed before they started law school in August, during the semester (M = 44 days later, SD = 2 days), during exams (M = 57 days later, SD = 3 days), after grades were released (M = 36 days later, SD = 4 days), and during interviews for summer internships (M = 29 days later, SD = 5 days). Stress due to law school was reported as lowest mid-semester and highest during finals (Roach et al., 2010). Variance due to people, occasions, and days was partitioned and used to calculate four kinds of reliability estimates (see Table 1; Cranford et al., 2006). Finally, we tested whether this variance could be predicted by daily assessment of mood and health behavior, effectively treating "error" variance as variance of interest (Segerstrom & Smith, 2012).

Method

Participants—Participants were 124 first-year law students. The sample was 55% female. Reported ethnicities were white (90%), Asian-American (1%), African-American (7%), and more than one race (2%). The mean age was 23.9 (SD = 2.9).

Procedure—All members of five consecutive incoming law school classes (2001–2005) received recruitment packets during the summer before starting law school. If interested in the study, they returned a signed informed consent, contact information, and a screening form for exclusion criteria related to mental health (e.g., self-reported history of impairment

of function for 2 weeks or more), physical health (e.g., autoimmune disease), and substance use (e.g., more than two drinks of alcohol every day). At each measurement occasion, participants collected saliva samples at home over three consecutive weekdays at waking, 30 minutes post-waking, noon, 5 pm, and 9 pm. Participants completed a health behavior and mood questionnaire on each of the sampling days. They received $50 for their participation at each time point.

Measures—*Affect* was measured daily in the evening using the PANAS-X with the "today" instruction (Watson & Clark, 1994). There were 4 subscales measuring negative affects (Fear, Hostility, Guilt, and Sadness), 3 subscales measuring positive affects (Joviality, Self-Assurance, and Attentiveness), and 4 measuring other affective states (Shyness, Fatigue, Serenity, and Surprise).

Health behaviors were measured daily in the evening with a series of items employing 11-item Likert scales ranging from 0 to 100, with anchors of "much less than usual" and "much more than usual". The items referred to daily "physical activity or exercise", "your health (how generally well you felt)?", "stressful", "pain", "sleep", "protein (e.g., meats or vegetable protein)/fat (e.g., oils or butter)/complex carbohydrate (e.g., starches, fruits)/sugar (e.g., sugared soft drinks, candy, or anything containing refined sugar)", "caffeinated beverages", and "alcoholic beverages". Individuals who did not drink caffeinated beverages or alcohol were instructed to leave those respective items blank.

Cortisol: Saliva samples were collected in pre-labeled Salivettes (Walter Sarstedt Inc., Newton, North Carolina), centrifuged, aliquoted, and frozen at −80 °C. Cortisol levels were assessed using an enzyme immunoassay (Salimetrics, Inc., State College, PA). Assay sensitivity was 0.007 ug/dL. The inter-assay CV was 7.76% for the low control and 4.66% for the high control. The intra-assay CV average was 6.50% using the low control and 3.19% using the high control.

There were 572 occasions of data collected of a possible 620 (124 enrollees*5 occasions; 92%). Most missing data were due to drop-outs (38; 6%). The rest were idiosyncratically due to missing both questionnaire data and cortisol (5; 1%) or only cortisol (5; 1%) at one occasion. Within occasions, each time point had 89–96% valid data. The primary reasons for missing data were missing samples, insufficient samples, and undetectable levels in the samples.

Data were excluded if reported collection time was outside a window defined as 4 SD earlier or later than mean collection time, in minutes (0.2% of cases were excluded). In cases in which subjects provided saliva but did not record the collection time for the sample (3% of cases), time values were imputed using the expectation-maximization algorithm in SPSS (cortisol values were not imputed). Of a total of 8,395 cortisol values, 11 were excluded on the basis that raw cortisol values were more than 4 standard deviations from the mean, calculated over collection time: High values were maintained in reassays performed after dilution, and no explanation for elevation could be found in the comments regarding sample collection.

The 30 minute post-waking sample was not used in the present study. The physiological, psychological, and clinical relevance of the cortisol awakening response (CAR) differs from that of other measures of salivary cortisol that integrate secretion level over the entire waking period; furthermore, the CAR and overall diurnal secretion are under different neurobiological control systems (Kraemer et al., 2006; Kumari et al., 2011; Sephton et al., 2000, 2013).

The distribution of raw cortisol values is typically skewed and the normal diurnal profile may be approximated by an exponential curve, so raw values were log transformed. Log cortisol values were regressed on the exact, reported time of sample collection. As per recommendations (Kramer et al., 2006), the diurnal slope calculation was anchored to waking cortisol, excluding the 30-minute post-waking sample. The unstandardized beta weight measured cortisol slope. Flatter diurnal cortisol slopes may reflect an overall flattening of the HPA rhythm, phase changes in circadian rhythm, ultradian rhythms, elevated trough levels, or blunted peak levels. Abnormally timed peaks occur in cases of phase change or ultradian rhythm (cycles shorter than 24 hours); in which peaks may occur in the afternoon, evening, or nighttime hours rather than in the morning (Sephton et al., 2000).

Mean cortisol comprised the mean of the log cortisol values. The area under the curve (AUC) calculation used mean raw cortisol values (C_x) and collection times (T_x) calculated over three days of assessment for each suggested time of collection (x = wake, noon, 5pm, or 9 pm). AUC was calculated by trapezoidal estimation using the formula: AUC = . $5[(C_{wake}+C_{noon})x(T_{noon}-T_{wake})] + .5[(C_{noon}+C_{5pm})x(T_{5pm}-T_{noon})]+.5[(C_{5pm}+C_{9pm})x(T_{9pm}-T_{5pm})]$

Data analysis

Generalizability and decision analyses: In the present study, the completely crossed *facets* of the G study were people, occasions, and days. SAS PROC GLM was used to obtain mean squares values. These values were used to calculate the generalizability coefficient for each facet and interaction (Brennan, 2001). The percentage of variance attributable to each term was calculated by dividing each coefficient by the total. Estimates from the G study were then used in the D study to estimate four kinds of reliability (Table 1). See Cranford and colleagues (2006) for a discussion of how the equations are constructed. Briefly, the equations are as follows, where p = person; o = occasion; d = day; m = number of days per occasion; and n = number of occasions:

Design I: $(\sigma^2_p+\sigma^2_{pd}/m)/(\sigma^2_p+\sigma^2_{pd}/m+\sigma^2_{pod}/m)$

Design II: $(\sigma^2_p+\sigma^2_{pd}/m)/(\sigma^2_p+\sigma^2_{pd}/m+\sigma^2_o+\sigma^2_{po}+\sigma^2_{pod}/m)$

Design III: $(\sigma^2_p+\sigma^2_{pd}/m)/(\sigma^2_p+\sigma^2_{pd}/m+\sigma^2_{pod}/mn)$

Design IV: $(\sigma^2_{po})/(\sigma^2_{po}+\sigma^2_{pod}/m)$

Within-day decisions: Following precedent (Kraemer et al., 2006), we calculated the cortisol slope for each day using permutations of sample combinations and used pairwise correlations to examine the degree to which each permutation correlated with the same slope

calculated with all available samples. Because each observation was not independent, we first calculated the correlation across observations within each person having 3 or more occasions of observation (N = 115), converted these 115 mean correlations to Fisher's z', took their mean across people, and converted the mean z' back to *r* to report the mean correlation, across occasions and people, between that permutation and the slope with all available samples.

Multi-level models: To explore sources of the variability identified in the generalizability study, multilevel models with health behaviors and mood as predictors and cortisol parameters as outcomes were tested using SAS PROC MIXED with maximum likelihood estimation. The model had days at Level l, occasions at Level 2, and people at Level 3. To identify the level at which each predictor accounted for cortisol variance, each was partitioned into three terms. The first term was the grand mean for each person across occasions and days. This term could account for differences between people in cortisol. The second term was the deviation of the occasion-level mean from the grand mean. This term could account for differences at different occasions within people. Finally, the third term was the deviation of the day-level from the occasion mean. This term could account for differences on different days within occasions.

For each predictor, the change in the −2 log likelihood of the model after adding all three terms tested whether that predictor accounted for statistically significant variance in cortisol, using the χ^2 distribution. To reduce the possibility of Type I error, a familywise Bonferroni correction was applied to the analyses of the health behavior predictors (.05/11 = .005) and the affect predictors (.05/11 = .005). If there was significant prediction, the individual terms were examined to determine at which level prediction occurred. This result is reported using the gamma weight, which is analogous to an unstandardized beta weight in regression.

Results

Generalizability Study: Variability—For all three cortisol parameters (see Table 2), the largest amount of variance was attributable to idiosyncratic interactions between person, occasion, and day (i.e., effects specific to a particular person at a particular occasion on a particular day), plus error (the two are statistically indistinguishable). For slope, approximately 10% of variance was attributable to stable differences between people. This value is unacceptably low if cortisol slope on a single day is used as a measure to estimate differences between people. This estimate was higher for diurnal mean and AUC, at approximately .24 and .19, although these estimates also indicate that a single day's measurement yields unacceptably low values to estimate differences between people.

There were meaningful amounts of variance attributable to individual differences in the effects of occasion. Despite negligible amounts of variance due to occasion *per se*, there was evidence that individuals reacted to occasions in idiosyncratic ways. Finally, there were no systematic differences between days (e.g., Day 1 was not systematically different – higher or lower – from Day 2 or Day 3, nor Day 2 from Day 3) or occasions (e.g., the first occasion was not systematically different – higher or lower – from the other occasions). The last line of Table 2 shows that, in comparison with mean and AUC, diurnal slope has the least

systematic variance when sources related to the person, occasion, and day are all considered. That is, diurnal slope has the most variance due to effects specific to a particular day at a particular occasion experienced by a particular person, as well as error.

Decision Study: Reliability—Although there is no systematic effect of day or occasion, there is substantial variance in the person*occasion*day term. If one wishes to "wash out" the day portion to get a reliable estimate of a person's cortisol profile when measured at a single occasion, or to "wash out" the day and occasion portion to get a reliable estimate of a person's cortisol profile when measured across occasions, then one must assess cortisol across a sufficient number of days. The number of days required to do so is estimated in the decision study. Table 3 shows the results of the decision study estimating reliability of measurement between people at the same occasion; reliability of measurement between people at different occasions (separated by 4–8 weeks); and reliability of measurement between occasions within people. The results from the present design, with 3 days collected at each occasion, are in bold. The Table also shows predicted reliabilities if data were collected for more days.

Using a standard of adequate reliability (.60), the 3 days of data adequately characterized differences between people at a single occasion for mean cortisol, and only one additional day would be required to characterize differences between people in AUC with adequate reliability. When the specific numbers of days projected to reach adequate reliability of measurement was computed, 11 or more days would adequately characterize differences between people at a single occasion for diurnal slope. Only mean cortisol measurement would reach high reliability (.80) at 7 days; AUC would reach this standard at 11 days, and the slope would not reach it even at 21 days. These estimates assume measurement at the same occasion. The next set of estimates illustrates the decrease in reliability of measurement resulting from assessment at different occasions. Under this assumption, none of the parameters characterized differences between people with adequate reliability.

The third set of estimates shows estimated measure reliability over 3 days of cortisol as they characterize differences between occasions within people. These estimates show adequate reliability for measures of mean cortisol and AUC. The diurnal slope would require 8 days of collection at each occasion for the measurement of change over occasions within people to be reliable. Mean cortisol measurement would reach high reliability (.80) at 5 days; AUC would reach this standard at 8 days, and the slope would reach it with 21 days of collection.

Table 4 shows reliability estimates for measurement of cortisol parameters across 3 days and across all 5 occasions (i.e., 15 total days) to characterize differences between people. Measures of all parameters had adequate between-person reliability, and mean cortisol and AUC were measured with good reliability. The Table also shows hypothetical measurement reliabilities if data were collected over more days and occasions. An important characteristic of these results is that they are symmetric: Adding more days to each occasion was not more or less effective in improving measurement reliability than adding more occasions. Rather, the total number of days collected appeared to be most important. For an adequately reliable measurement of diurnal slope, this total should be near 15 days; for mean, 3 days; and for AUC, 4 days.

How many samples per day for diurnal slope?—Another design decision concerns the number of samples per day required to accurately reproduce a diurnal slope calculated with more samples. Table 5 shows that all slope permutations that included the waking and 9 pm times faithfully reproduced the 4-point slope (r = .97 – .99).

Predictors of variability in cortisol—Does the variability in cortisol parameters identified in the generalizability analysis have systematic predictors related to affect or health behavior? There were two significant predictors of variability in the diurnal cortisol slope. First, adding pain predictors resulted in a significant model improvement ($\chi^2(3)=14.4$, p <002). Of the three terms, differences in pain across occasions was the best predictor of differences in diurnal slope, γ=07, SE=03, t(1520)=2.70, p <007. During measurement occasions when people reported more pain, diurnal cortisol slope was flatter. Second, adding fatigue predictors resulted in a significant model improvement ($\chi^2(3)=12.7$, p = 005). Of the three terms, differences in fatigue across days was the best predictor of diurnal slope, γ=94, SE=.31, t(1519)=3.01, p <003. On days when people reported more fatigue, diurnal cortisol slope was flatter. There were no significant predictors of diurnal mean or AUC.

Discussion

Recommendations vary regarding the number of days required to characterize diurnal cortisol parameters, particularly cortisol slope for comparisons of data between people, with adequate reliability of measurement. Whereas the MacArthur network suggested 3–4 days for mean cortisol and 6 or more days for cortisol slope, Kraemer and colleagues (2006) suggested that fewer days would be adequate, and Hruschka and colleagues (2005) recommended more. The results of this study among young adults are most in accordance with the recommendation of the MacArthur network. Although 3 days reached a minimal level of reliable measurement for the cortisol mean, and 4 days for AUC, 11 days would be necessary to estimate cortisol slope with adequate reliability to assess comparisons between people measured on similar occasions.

Because of the variance associated with individual differences in reaction to occasions (i.e., person*occasion variance: 13.9% for slope, 32.1% for mean, and 26% for AUC), designs in which people are measured on different occasions make it difficult to isolate person variance. Multiple weeks of collection would be necessary to do so, which might be prohibitive for participants as well as researchers.

This is also the first study to estimate how many days would be needed to differentiate different occasions within a single person, that is, reliable within-person measurement. These results resembled the estimates for reliable measurement between people. In order to characterize changes across occasions with adequately reliable assessment, at least 3 days of data collection per occasion were needed for mean cortisol, 4 days for AUC, and 8 days for diurnal slope.

However, daily variability is not necessarily something to be eliminated. By measuring predictors at the day level, “error” variance (i.e., the differences in cortisol levels or slopes across days) from the perspective of individual differences and changes over occasions became variance of interest. Note that the larger amount of “idiosyncratic” variability in

diurnal slope compared with mean and AUC may have *facilitated* its prediction by these daily predictors. A measure with little daily variance would provide little to be predicted.

Study 2

One possibility is that the variance in diurnal cortisol is differently distributed in different samples. Study 2 employed a longitudinal study of healthy older adults.

Method

Participants—Participants were 148 older adults. The sample was 58% female (N = 87). Reported ethnicities were white (96%, N = 142) and African-American (4%, N = 6). The mean age at the first occasion of data collection was 74.0 (SD = 6.1).

Procedure—Older adults (≥ 60 years) were recruited from clinics and the volunteer subject pool of the Sanders-Brown Center on Aging between July, 2006 and December, 2007. Interested individuals were contacted by phone and screened for exclusion criteria related to physical health: diseases or disorders affecting the immune system or chemotherapy or radiation treatment within the past 5 years, unwillingness to undergo vaccination or venipuncture, taking immunomodulatory medications including opiates and steroids, and taking more than two of particular classes of medications (psychotropics, antihypertensives, hormone replacement, or thyroid supplements).

Eligible participants were assessed every six months (M interval between occasions = 190 days, SD = 29 days). At each time point, participants were mailed materials for saliva collection and a health behaviors questionnaire as described for Study 1 (daily affect was not measured in Study 2). They received $20 for their participation at each time point.

Measures—*Health behaviors* were measured daily as in Study 1.

Cortisol: Salivary cortisol data collection procedures were similar to those described above, except that there were 3 samples per day: waking, 5 pm, and 9 pm. Assay sensitivity was 0.003 ug/dL. Data from the low and high control samples yielded respectively, average inter-assay CVs of 9.40% and 5.49% and intra-assay CVs of 5.31% and 3.13%. There were 417 occasions of data collected of a possible 592 (148 enrollees*4 occasions; 70%). The largest proportion of participants (36%) completed all 4 occasions; 20%, 3, 34%, 2, and only 10%, 1. All time points had > 98% valid data. Outliers and missing data were treated as in Study 1. Cortisol summary variables including diurnal mean and diurnal slope were calculated as in Study 1. The area under the curve (AUC) calculation used mean raw cortisol values (C_x) and collection times (T_x) calculated over three days of assessment for each suggested time of collection (x = wake, noon, 5pm, or 9 pm). AUC was calculated by trapezoidal estimation using the formula: $AUC = .5[(C_{wake}+C_{5pm})x(T_{5pm}-T_{wake})] + .5[(C_{5pm}+ C_{9pm})x(T_{9pm}-T_{5pm})]$

Data analysis—All analyses were performed as described for Study 1, except SAS PROC VARCOMP was used to obtain variance components.

Results

Generalizability Study: Variability—As in Study 1, for all cortisol parameters (see Table 2), the largest amount of variance was attributable to idiosyncratic interactions between person, occasion, and day, including error. For diurnal slope, approximately 10% of variance was attributable to differences between people and for AUC, 11%. This estimate was higher for the diurnal mean, at approximately 28%. These estimates are similar to the estimates in Study 1 (11%, 19%, and 24%, respectively), despite differences in samples and interval between occasions. Also as in Study 1, there were meaningful amounts of variance (20%–29%) attributable to individual differences in the effects of occasion, but no apparent effects of occasion *per se*. Finally, there was no meaningful variance attributable to systematic differences between days.

Decision Study: Reliability—Table 3 shows the results of the decision study estimating reliable measurement between people at the same, fixed occasion; between people at different occasions (separated by 6 months); and within people when they are measured across occasions. The results from the present design, with 3 days of data collected at each occasion, are in bold. The Table also shows the predicted measure reliabilities if data were collected for more days.

Again, the 3 days of data provided adequate reliability of measurement (.60) only for assessment of differences in mean cortisol between people at a single occasion. When the specific numbers of days projected to reach adequate measure reliability were computed, adequate reliability would be reached at 8 days for AUC, and 10 days for diurnal slope. Only mean cortisol would be assessed with high reliability (.80) at 6 days. AUC would reach the high reliability standard at 21 days, and slope at more than 21 days of collection. These estimates assume that everyone was measured at the same occasion. Measurement of none of the parameters achieved adequate reliability to discriminate between people if they were measured at different occasions, even at 100 days.

The third set of estimates shows the reliability of measurement for the 3 days of cortisol as they characterize differences between occasions within people. These estimates of within-person diurnal cortisol measurement reliability show adequate reliability only for mean cortisol, although AUC would require only one additional day to reach this threshold. Diurnal slope measurement would require 5 days of collection for adequate reliability as it changes within people. High reliability of measurement (.80) would require 6 days of collection for estimates of mean cortisol, 9 days for AUC, and 14 days for the diurnal slope.

Table 6 shows the reliability of measurement for the 3 days of cortisol across all 4 occasions (i.e., 12 total days) in the characterization of differences between people. All measures had adequate between-person reliability, and measurement of mean cortisol had high reliability. The Table also shows hypothetical measure reliabilities if data were collected over more days and more occasions. As in Study 1, these results were symmetric across days and occasions. For diurnal slope, the total number of days collected across occasions should be near 9 days; for mean, 3 days; and for AUC, 4 days.

How many samples per day for diurnal slope?—Table 5 shows that again the permutation of the slopes using the waking and 9 pm time points very faithfully reproduced the 3-point (waking, 5 pm, and 9 pm) slope (r = .99).

Predictors of variability in cortisol—None of the health behaviors reached statistical significance in predicting variability in mean cortisol, AUC, or diurnal slope (all $\chi^2(3) \leq 10.0$, all $p > .01$).

Discussion

The results from Study 2 paralleled those of Study 1. The minimum numbers of days needed to construct a reliable measure of between-person variance were 3 for mean cortisol, 8 for AUC, and 10 for diurnal slope, using a liberal estimate of adequate reliability of .60. Although this was a demographically different sample with longer periods between occasions than Study 1, the amounts of variance due to person, occasion, and day were similar and add confidence in the design recommendations.

General Discussion

Two longitudinal studies generated estimates of the procedures that would best allow researchers to estimate between-and within-person differences in salivary cortisol with adequate reliability. Across samples, most of the variance in diurnal cortisol slopes stemmed from person (individual differences; 11%), person by occasion (occasion effects differing across people; 14–20%), and person by day by occasion (effects specific to a particular person at a particular occasion on a particular day and measurement error; 68–75%) effects. The variance in mean cortisol concentrations also stemmed from person (24–28%), person by occasion (29–32%), and person by day by occasion (40–41%) effects. AUC results were similar to those for mean cortisol (person, 19–35%; person by occasion, 26–28%; person by day by occasion, 53–60%). Although previous studies have not been able to isolate all of these sources of variance, the proportion of person variance is in line with other estimates when the interval between assessments was weeks to months (Kirschbaum et al., 1990; Hruschka et al., 2005; Rotenberg et al., 2012; Shirtcliff et al., 2012). Compared with other outcomes, the proportions of person variance in diurnal cortisol parameters are similar to several dimensions of daily mood such as anxiety and anger, although lower than others such as fatigue and vigor, as well as positive and negative affect measured broadly (Cranford et al., 2006; Röcke, Li, & Smith, 2009). They are substantially lower than those of many immunological parameters, some of which reflect over 50% variance due to person effects (i.e., stable individual differences; Segerstrom & Smith, 2012, Table A). They also differ in some respects from the proportions of person variance in salivary alpha amylase (sAA) measured in the present Study 1, with the sAA estimate for the diurnal slope only slightly higher, but those for mean and AUC substantially higher (Out, Granger, Sephton, & Segerstrom, 2013).

Decision analyses informed the number of days of data collection needed to achieve minimum reliability of measurement (.60). Different study designs rely on different estimates of reliability. For studies comparing differences in measures of diurnal cortisol between people at a single occasion, how many days of data collection are needed to reliably

measure these differences? Our analyses suggest roughly a minimum of 10 days for diurnal slope, 4–8 days for AUC (depending on the sample), and 3 days for mean cortisol. For studies comparing differences in measures of diurnal cortisol between people at different occasions, none of the parameters described above can be reliably measured even with over 21 days of data, thereby making it an ineffective design. Finally, for longitudinal studies interested in comparing within-person differences in measures of diurnal cortisol across occasions, we suggest collecting a minimum of 5–10 days at each occasion for diurnal slope and 3 days for mean cortisol and AUC. To our knowledge, these are the first estimates for within-subject longitudinal designs. It is worth noting that the designs in our studies involved intervals of weeks to months between measurement occasions, which may have increased our ability to detect within-person change across occasions by allowing enough time for change to occur.

It may appear that the number of days necessary to reliably assess cortisol makes longitudinal designs with multiple occasions cost-prohibitive. However, cortisol slopes calculated using two samples per day (waking and evening) strongly correlated with estimates using several samples per day (see also Kraemer et al., 2006). A recommended strategy would be to collect fewer samples per day over more days. For instance, a protocol in which participants provide 2 samples over the course of 10 days would be better able to detect between-and within-person differences in cortisol slope than one in which participants provide 5 samples a day over 4 days. It maximizes power to capture differences in diurnal slope while keeping the cost equivalent. It is also likely less burdensome for the participant, as it requires fewer interruptions to his or her daily routine.

Although increasing the number of data collection days has a financial cost, there are also costs to not doing so. Collecting insufficient data will produce sub-optimal reliability of measurement and large error variance, which will in turn increase the probability of Type II errors. In the case of a Type II error caused by insufficient reliability, *all* of the financial cost has gone to naught. Costs of unreliability therefore include both postponement to scientific advancement and expenditures of researcher and participant resources (Halpern et al., 2012; Segerstrom & Smith, 2012). A related issue concerns replicability, a concern that has re-emerged recently in psychology and other fields. Two studies are most likely to converge on the same finding when their estimates are both accurate, and this is less likely to be true when their measures are unreliable (Segerstrom et al., 2006).

The greatest variability was due to idiosyncratic interactions among person, occasion, and day. Although such variability is usually treated as error, there are likely unmeasured variables that partially account for this variance. Researchers could therefore transform unexplained error into “good” variance (Segerstrom & Smith, 2012). In the present study, for example, fatigue and pain were related to “error” variance in cortisol slope, manifest as flattening of the diurnal cortisol profile. In turn, flattened diurnal cortisol slopes are clinically relevant, both in terms of psychological (Carlson et al., 2007; Cohen et al., 2006; Miller et al., 2007) and physical health (Sephton et al., 2000; Kumari et al., 2011; Sephton et al., 2013). In addition to mood and health behavior, other daily measures may provide insight into the causes of day-to-day variation in diurnal cortisol.

These results can also inform covariate design. Covariates should account for the "error" variance associated with facets *not* of interest, thereby increasing the proportion of variance in the facet of interest. For example, where differences between people are of interest (i.e., the Person facet), covariates should be selected that can account for differences due to days and occasions (i.e., facets including Occasion and Day effects). For example, in Study 1, the same person-level reliability of diurnal slope measurement could be expected with 1 less day of measurement if "error" variance due to wave- and day-level fatigue and pain was removed. It is important to note that although appropriately used covariates may improve "physiometric" properties, they can compromise the ability to compare results across studies that use different covariates and as such have different (residualized) outcome variables (Segerstrom, 2009).

These studies are not without limitations. First, both used relatively healthy samples. Although the results closely corroborated one another, further research should test how well results generalize to other populations, including those with medical illness. The generalizability of cortisol parameters may differ across populations; for example, in the present studies, variance due to stable individual differences in AUC was lesser in older than in younger adults (Table 1), leading to different projections in the decision study (Table 2). Second, the saliva collection procedure was not monitored and time of day was self-reported. Although this procedure reduced cost and burden and may be informative in terms of design decisions for similar studies, it also introduced variability. Across data from both studies, the waking mean time was 7:59 am, with a relatively large interquartile range (2:00). This variability is to be expected, considering that time of waking was controlled by the participant and not by the study design. Other time points showed good compliance and less variability, with the noon sample collected on average at 12:32 pm (0:50), the 5 pm at 5:28 pm (0:45) and the 9 pm at 9:32 pm (0:55). Further research is needed to determine how effective more standardized and highly monitored procedures would be in improving the reliability of diurnal cortisol estimates.

Despite these potential limitations, these studies provide insights into future directions in research employing diurnal cortisol. Many salivary cortisol studies employ between-subject designs. There is less longitudinal, within-subject research exploring how cortisol parameters change within people over time. Future research should also begin modeling within-person effects, and doing so with sufficient number of data collection days.

Acknowledgments

FUNDING ACKNOWLEDGEMENTS

The research reported here was supported by the Dana Foundation and the National Institutes of Health MH61531-R01, AG026307-R01, AG033629-K02, and AG028383-P30.

References

Brennan, RL. Generalizability Theory. New York: Springer; 2001.

Carlson LE, Campbell TS, Garland SN, Grossman P. Associations among salivary cortisol, melatonin, catecholamines, sleep quality and stress in women with breast cancer and healthy controls. J. Behav. Med. 2007; 30:45–58. [PubMed: 17245618]

Cohen S, Schwartz JE, Epe lE, Kirschbaum C, Sidney S, Seeman T. Socioeconomic status, race, and diurnal cortisol decline in the Coronary Artery Risk Development in Young Adults (CARDIA) Study. Psychosom. Med. 2006; 68:41–50. [PubMed: 16449410]

Cranford JA, Shrout PE, Iida M, Rafael iE, Yip T, Bolger N. A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? Pers. Soc. Psychol. Bull. 2006; 32:917–939. [PubMed: 16738025]

Estabrook R, Grimm KJ, Bowles RP. A Monte Carlo simulation study of the reliability of intraindividual variability. Psychol. Aging. 2012; 27:560–576. [PubMed: 22268793]

Fisher PA, Stoolmiller M, Gunnar MR, Burraston BO. Effects of a therapeutic intervention for foster preschoolers on diurnal cortisol activity. Psychoneuroendocrinol. 2007; 32:892–905.

Gex-Fabry M, Jermann F, Kosel M, Rossier MF, Van der Linden M, Bertschy G, Bondolfi G, Aubry JM. Salivary cortisol profiles in patients remitted from recurrent depression: One-year follow-up of a mindfulness-based cognitive therapy trial. J. Psychiatr. Res. 2012; 46:80–86. [PubMed: 21982583]

Golden SH, Wand GS, Malhotra S, Kamel I, Horton K. Reliability of hypothalamic–pituitary–adrenal axis assessment methods for use in population-based studies. Eur. J. Epidemiol. 2011; 26:511–525. [PubMed: 21533585]

Goodyer IM, Park RJ, Netherton CM, Herbert J. Possible role of cortisol and dehydroepiandrosterone in human development and psychopathology. Brit. J. Psychiat. 2001; 179:243–249.

Halpern CT, Whitsel EA, Wagner B, Harris KM. Challenges of measuring diurnal cortisol concentrations in a large population-based field study. Psychoneuroendocrinol. 2012; 37:499–508.

Hruschka DJ, Kohrt BA, Worthman CM. Estimating between-and within-individual variation in cortisol levels using multilevel models. Psychoneuroendocrinol. 2005; 30:698–714.

Ice GH, Katz-Stein A, Himes J, Kane RL. Diurnal cycles of salivary cortisol in older adults. Psychoneuroendocrinol. 2004; 29:355–370.

Kertes DA, van Dulmen M. Latent state trait modeling of children's cortisol at two points of the diurnal cycle. Psychoneuroendocrinol. 2012; 37:249–255.

Kirschbaum C, Steyer R, Eid M, Patalla U, Schwenkmezger P, Hellhammer DH. Cortisol and behavior: 2. Application of a latent state-trait model to salivary cortisol. Psychoneuroendocrinol. 1990; 15:297–307.

Kraemer HC, Giese-Davis J, Yutsis M, O'Hara R, Neri E, Gallagher-Thompson D, Taylor B, Spiegel D. Design decisions to optimize reliability of daytime cortisol slopes in an older population. Am. J. Geriatr. Psychiatry. 2006; 14:325–333. [PubMed: 16582041]

Kudielka BM, Gierens A, Hellhammer DH, Wüs tS, Schlotz W. Salivary cortisol in ambulatory assessment—some dos, some don'ts, and some open questions. Psychosom. Med. 2012; 74:418–431. [PubMed: 22582339]

Kumari M, Shipley M, Stafford M, Kivimaki M. Association of diurnal patterns in salivary cortisol with all-cause and cardiovascular mortality: findings from the Whitehall II study. J. Clin. Endocrinol. Metab. 2011; 96:1478–1485. [PubMed: 21346074]

Lupien S, Lecours AR, Schwartz G, Sharma S, Hauger RL, Meaney MJ, Nair NPV. Longitudinal study of basal cortisol levels in healthy elderly subjects: Evidence for subgroups. Neurobiol. Aging. 1996; 17:95–105. [PubMed: 8786810]

MacArthur Network. Salivary cortisol measurement. Retrieved from http://www.macses.ucsf.edu/research/allostatic/salivarycort.php

Miller GE, Chen E, Zhou ES. If it goes up, must it come down? Chronic stress and the hypothalamic-pituitary-adrenocortical axis in humans. Psychol. Bull. 2007; 133:25–45. [PubMed: 17201569]

Out D, Granger DA, Sephton SE, Segerstrom SC. Disentangling sources of individual differences in diurnal salivary a-amylase: Reliability, stability, and sensitivity to context. Psychoneuroendocrinol. 2013; 38:367–375.

Ranjit N, Diez-Roux AV, Sanchez B, Seeman T, Shea S, Shrager S, Watson K. Association of salivary cortisol circadian pattern with cynical hostility: Multi-ethnic study of atherosclerosis. Psychosom. Med. 2009; 71:748–755. [PubMed: 19592518]

Roach AR, Salt CE, Segerstrom SC. Generalizability of repetitive thought: Examining stability in thought content and process. Cogn. Ther. Res. 2010; 34:144–158.

Röcke C, Li SC, Smith J. Intraindividual variability in positive and negative affect over 45 days: Do older adults fluctuate less than young adults? Psychol. Aging. 2009; 24:863–878. [PubMed: 20025402]

Rotenberg S, McGrath JJ, Roy-Gagnon MH, Tu MT. Stability of the diurnal cortisol profile in children and adolescents. Psychoneuroendocrinol. 2012; 37:1981–1989.

Segerstrom SC. Biobehavioral controls: Threats to psychoneuroimmunology research? Brain Behav. Immun. 2009; 23:885–886. [PubMed: 19463944]

Segerstrom SC, Lubach GR, Coe CL. Identifying immune traits and biobehavioral correlates: Generalizability and reliability of immune responses in rhesus macaques. Brain Behav. Immun. 2006; 20:349–358. [PubMed: 16293393]

Segerstrom, SC.; Smith, GT. Methods, variance, and error in psychoneuroimmunology research: The good, the bad, and the ugly. In: Segerstrom, S., editor. Oxford Handbook of Psychoneuroimmunology. New York: Oxford U Press; 2012. p. 421-432.

Sephton SE, Lush E, Dedert EA, Floyd AR, Rebholz WN, Dhabhar FS, Spiegel D, Salmon P. Diurnal cortisol rhythm as a predictor of lung cancer survival. Brain Behav. Immun. 2013; 30:S163–S170. [PubMed: 22884416]

Sephton SE, Sapolsky RM, Kraemer HC, Spiegel D. Diurnal cortisol rhythm as a predictor of breast cancer survival. J. Nat. Cancer Inst. 2000; 92:994–1000. [PubMed: 10861311]

Shavelson, RJ.; Webb, NM. Generalizability Theory: A Primer. Newbury Park: Sage; 1991.

Shirtcliff EA, Allison AL, Armstrong JM, Slattery MJ, Kalin NH, Essex MJ. Longitudinal stability and developmental properties of salivary cortisol levels and circadian rhythms from childhood to adolescence. Dev. Psychobiol. 2012; 54:493–502. [PubMed: 21953537]

Watson, D.; Clark, LA. The PANAS-X: Manual for the Positive and Negative Affect Schedule-expanded form. Iowa City: U of Iowa; 1994. Unpublished manuscript

Highlights

- Two studies characterized sources of variability (day, occasion, and person) in diurnal slope, AUC, and diurnal mean of salivary cortisol and determined how many days of measurement would be needed to construct reliable measures for the detection of between-person and within-person differences.
- Three days of measurement provided the minimal level of reliability in mean cortisol to detect between-person and within-person differences.
- Three days of measurement provided the minimal level of reliability in AUC for within-person differences; 4–8 days were necessary for between-person differences.
- Five to 8 days of measurement provided the minimal level of reliability in diurnal slope; 10 days were necessary for between-person differences.

Highlights

- Two studies characterized variability and reliability in diurnal mean, AUC, and diurnal slope of salivary cortisol.
- Reliability estimates were provided for between- and within-person designs.
- Between and within people, 3 measurement days provided minimal reliability of mean cortisol.
- Within people, 3–4 measurement days provided minimal reliability of AUC; between people, 4–8 days.
- Within people, 5–8 measurement days would provide minimal reliability of diurnal slope; between people, 10–11 days.

Table 1

Four kinds of reliability (Cranford et al., 2006), with descriptions and design examples

	Description	Design Decision Question	Empirical Example
I. Between subjects at the same occasion of measurement	**Cross-sectional, single occasion**	How many days of data are required to construct a reliable measure of differences between people measured at the same occasion?	Finding: Education correlated negatively with diurnal cortisol slope (Cohen et al., 2006). Sampling: 5 samples per day, 1 day, 781 participants
II. Between subjects at different occasions of measurement	**Cross-sectional, different occasions**	How many days of data are required to construct a reliable measure of between people measured at different occasions?	Finding: Psychological and sleep variables measured once were unrelated to diurnal cortisol mean, slope, or AUC measured once in breast cancer patients at varied points post-diagnosis (Carlson et al., 2007) Sampling: 4 samples per day, 1 day, 33 participants
III. Between subjects across multiple occasions	**Longitudinal or repeated measures, multiple occasions**	How many days of data over how many time points are required construct a reliable measure of stable differences between people?	Finding: A personality subscale of the SCL-90 measured at one time point significantly correlated with mean basal cortisol levels averaged across 3–6 years (Lupien et al., 1996) Sampling: 24 samples per day, 1 day, 3–6 occasions, 19 participants
IV. Within subjects, change across occasions	**Longitudinal or repeated measures, multiple occasions**	How many days of data at each time point are required construct a reliable measure of differences between occasions within the same person?	Finding: Foster children receiving an intervention had steeper diurnal cortisol slopes over time; children in regular foster care had flatter slopes over time (Fisher et al., 2007). Sampling: 2 samples per day, 2 days, 12 occasions, 117 participants Note: A latent variable was used to isolate occasion from day variance

Note: See Kudielka and colleagues (2012) for design considerations regarding within-day research questions such as those associated with ecological momentary assessment

Table 2

Generalizability study results for cortisol parameters

		Study 1			Study 2		
	Variance due to:	**Diurnal slope**	**Mean**	**AUC**	**Diurnal slope**	**Mean**	**AUC**
Person	Differences between people	10.6%	23.5%	19.2%	10.8%	27.6%	11.4%
Occasion	Differences between occasions	0.3%	2.3%	0.8%	0%*	2.5%	0.5%
Day	Differences between days	0.0%	1.1%	0%*	0%*	0.3%	0%*
Person*Occasion	Occasion effects different for different people	13.9%	32.1%	26.0%	20.0%	28.6%	27.5%
Person*Day	Day effects different for different people	0.5%	0%*	1.3%	0.0%*	0.6%	0%*
Occasion*Day	Day effects different for different occasions	0%*	0.1%	0%*	1.8%	0.6%	0.2%
Person*Occasion*Day, Error	Idiosyncratic differences and error	74.7%	41.1%`	52.7%	67.5%	40.1%	60.4%

*Small negative variance estimate set to 0.

Note: AUC = area under the curve

Table 3

Results of decision study reliability estimation for diurnal slope, diurnal (log) mean, and waking slope in Study 1 and Study 2. See Table 1 for description of the designs.

	Study 1			Study 2		
	Slope	**Mean**	**AUC**	**Slope**	**Mean**	**AUC**
Design I. Between subjects at the same occasion of measurement						
Number of days						
3	**.30**	**.63**	**.53**	**.32**	**.67**	**.36**
7	.50	.80	.72	.53	.83	.57
10	.59	.85	.79	.61	.87	.65
14	.67	.89	.84	.69	.91	.73
21	.75	.92	.88	.77	.94	.80
Design II. Between subjects at different occasions of measurement						
Number of days						
3	**.22**	**.31**	**.31**	**.20**	**.38**	**.19**
7	.30	.34	.36	.27	.43	.24
10	.33	.35	.38	.29	.44	.25
14	.35	.36	.39	.30	.45	.26
21	.37	.36	.40	.32	.46	.27
Design IV. Within subjects, change across occasions						
Number of days						
3	**.36**	**.74**	**.60**	**.47**	**.68**	**.58**
7	.57	.87	.78	.68	.83	.76
10	.65	.90	.83	.75	.88	.82
14	.72	.93	.87	.81	.91	.86
21	.80	.95	.91	.86	.94	.91

Note: AUC = area under the curve

Table 4

Results of decision study reliability estimation for diurnal slope, diurnal (log) mean, and area under the curve between people, across all occasions in Study 1. See Table 1 for description of the design.

	Diurnal slope				Mean				AUC			
Design III. Between subjects across multiple occasions												
N occasions	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**
N days												
1	.13	.30	.43	.60	.36	.63	.74	.85	.28	.54	.66	.80
3	.30	.56	**.68**	.81	.65	.84	**.89**	.95	.53	.77	**.85**	.92
5	.42	.68	.78	.88	.75	.89	.93	.97	.65	.85	.90	.95
10	.59	.81	.88	.93	.85	.94	.97	.98	.79	.92	.95	.97

Note: AUC = area under the curve

Table 5

Correlations of 2- and 3-point cortisol slopes with the 4-point cortisol slope as the standard in Study 1 (N = 115) and correlations of 2-point slopes with the 3-point slope in Study 2 (N = 84)

Points included	Mean correlation	Median correlation (25^{th}–75^{th} percentile)
Wake, noon, 5 pm	.86	.83 (.60 – .94)
Wake, noon, 9 pm	.99	.98 (.95 – .99)
Wake, 5 pm, 9 pm	.99	.99 (.96 – .99)
Wake, noon	.62	.62 (.16 – .83)
Wake, 5 pm	.84	.81 (.56 – .94)
Wake, 9 pm	.97	.97 (.92 – .99)
Wake, 5 pm	.80	.77 (.33–.94)
Wake, 9 pm	.99	.99 (.96–.99)

Table 6

Results of decision study reliability estimation for diurnal slope and diurnal (log) mean between people, across all occasions in Study 2. See Table 1 for description of the design.

	Diurnal slope					**Mean**					**AUC**				
Design III. Between subjects across multiple occasions															
N occasions	**1**	**3**	**4**	**5**	**10**	**1**	**3**	**4**	**5**	**10**	**1**	**3**	**4**	**5**	**10**
N days															
1	.14	.32	.39	.44	.61	.41	.68	.74	.78	.88	.16	.36	.43	.49	.65
3	.32	.59	**.66**	.71	.83	.67	.86	**.89**	.91	.95	.36	.63	**.69**	.74	.85
5	.44	.71	.76	.80	.89	.78	.91	.93	.95	.97	.49	.74	.79	.83	.90
10	.61	.83	.86	.89	.94	.87	.95	.96	.97	.99	.65	.85	.88	.90	.95

Note: AUC = area under the curve